

PAPER • OPEN ACCESS

Modeling and forecasting the parameters of a railroad transport system

To cite this article: O I Rebrin *et al* 2020 *J. Phys.: Conf. Ser.* **1679** 032018

View the [article online](#) for updates and enhancements.

 <p>The Electrochemical Society Advancing solid state & electrochemical science & technology 2021 Virtual Education</p> <p>Fundamentals of Electrochemistry: Basic Theory and Kinetic Methods Instructed by: Dr. James Noël Sun, Sept 19 & Mon, Sept 20 at 12h–15h ET</p> <p>Register early and save!</p>	
---	--

Modeling and forecasting the parameters of a railroad transport system

O I Rebrin, L A Zakharov, L A Derksen and V I Eremenko

Ural Federal University named after the first President of Russia B. N. Yeltsin. 19 Mira street, Ekaterinburg, 620002, Russia

Abstract. The article describes an algorithm for constructing a mathematical model to predict the demand for railway passenger transportation. The model based on the SARIMAX methodology, considering the influence of seasonality and external variables on demand. The average absolute error of prediction result of the created model with the calculated parameters is 8%. The created model can be used as a decision support system in the field of passenger transportation.

1. Introduction

One of the important elements of the success of a transport company is the correct forecasting of future demand for transport services. The data obtained as a result of forecasting makes it possible to more accurately plan operational activities, marketing and financial policies of the enterprise aimed at maximizing profits and reducing costs, increasing competitiveness. At the same time, demand is influenced by various external and internal factors, which complicates forecasting. Considering the influence of such factors is possible in complex models that require a lot of computing power to select the parameters. The aim of the research is to develop a mathematical model that considers the influence of seasonality and external factors to predict the demand for passenger transportation services of transport company.

Currently used prediction methods can be divided into two categories: parametric and nonparametric methods [1]. The main difference between these two categories is the functional relationship between the dependent and independent variables. Examples of nonparametric methods: neural networks [2,3], k-nearest neighbors [4,5], support vector machine [6,7], parametric - the family of autoregressive integrated moving average (ARIMA) methods [8–10], filters Kalman [11] and the method of maximum likelihood [12]. Compared to parametric methods, nonparametric methods are more efficient at predicting. However, nonparametric methods require a large amount of historical data and training, which makes parametric methods preferable for solving a number of applied problems.

The presented study uses data on the monthly passenger traffic of the intercity segment for the period from 2016 to 2019 provided by the Federal State Statistics Service of the Russian Federation.

2. Methods

The object of modeling is the transport system of railway passenger transportation. The investigated transport system can be described in two aspects: logistics and market. The market aspect reflects the traffic network from point to point, passenger and traffic flows along this network and is described by the matrix M (table 1). The logistic aspect reflects the railway network, passenger and transport (carriage) flows along this network, including those grouped into trains, which is described by the matrix



L (table 2). Figure 1 describes the diagram of passenger flows in the market and logistics aspects in the form of a directed acyclic graph, where vertices A, B, C and D denote railway stations, arcs denote passenger flows with a quantitative designation of demand.

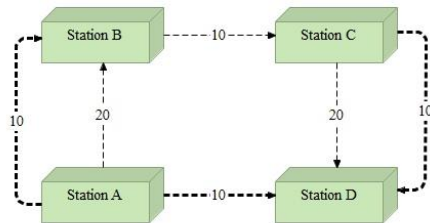


Figure 1. An example of describing passenger flows in the transport system in terms of logistic and market aspects. The numbers represent the quantity of tickets purchased (demand), the thin arrows represent the logistics aspect represented in the L matrix, and the thick arrows the market aspect represented in the M matrix.

The matrices L and M in tables 1 and 2 describe the entire transport system. Matrices are specified for each date d on which streams from i to j are sent. The proposed format makes it possible to represent demand as a set of time series for each arc of the graph, which is convenient for analyzing and building a predictive model for any slice of passenger traffic.

Table 1. Matrix M describing passenger traffic in the market aspect.

	A	B	C	D
A	0	10	0	10
B	0	0	0	0
C	0	0	0	10
D	0	0	0	0

Table 2. Matrix L, describing passenger traffic in the logistic aspect.

	A	B	C	D
A	0	20	0	0
B	0	0	10	0
C	0	0	0	20
D	0	0	0	0

One of the common methods for forecasting time series related to the class of parametric is the method of modeling autoregressive moving average (ARIMA), which generalizes the autoregressive model (AR) and the moving average (MA) model. The ARIMA model is based on the key assumption that historical process performance is good predictor of future performance. Extending the model to SARIMAX also assumes that the current value is influenced by the “seasonality” (i.e. cyclicity) of the process, as well as the external observable values of other processes. As applied to our problem, the mentioned seasonality corresponds to the natural fluctuations in demand due to the seasons and days of the week. In this case, external observed values will be associated with the production calendar, which determines the holidays and their relative position, as well as their location within the week and season.

To implement forecasting demand for a quarter, an algorithm was developed, presented in the form of a archimate diagram in figure 2. The input data for the algorithm are historical data on ticket sales. The output is the demand forecast for the quarter as a time series.

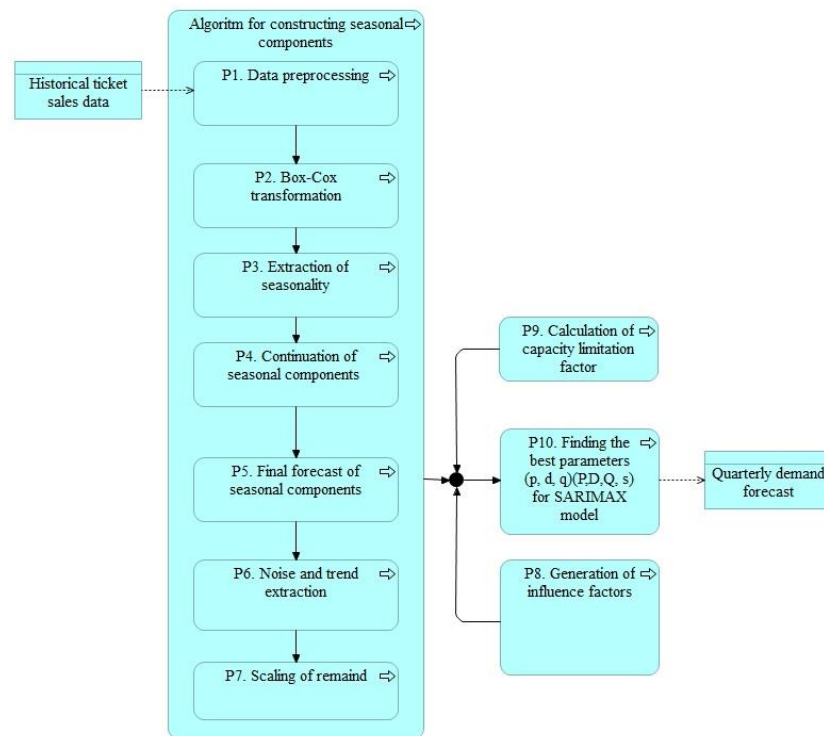


Figure 2. Archimate diagram of the forecasting algorithm. Each rectangle corresponds to the corresponding step of the prediction algorithm P1 – P9.

Below is the sequence of steps of the forecasting algorithm (figure 2) used to build the passenger traffic demand model. At stage P2 is performed Box-Cox transformation [13] to reduce the dispersion of the row:

$$y = \frac{(x\lambda - 1)}{\lambda}, \text{ if } \lambda \neq 0 \quad (1)$$

$$y = \ln(x), \text{ if } \lambda = 0 \quad (2)$$

Where x is initial time series as a vector, y is time series after conversion, λ is transformation parameter, which is selected from the condition of maximizing the logarithmic function of likelihood.

In step P3, seasonality is selected by STL decomposition [14] and consists of 3 subdivisions:

- selection of weekly seasonality in step 7;
- extraction of the annual seasonality in step 364 to get to the same day of the week;
- extraction of the annual seasonality in increments 365 to get on the same date.

At stage P4, seasonal components are continued for the length of the forecast.

In step P5 the final forecast of seasonal components is made in the form of the sum of weekly seasonality and the average between annual seasonality.

In step P6 the seasonal component is subtracted from the initial series to get the remainder in the form of noise and trend, which will be predicted.

In step P7, the residue is scaled using the classical data rationing method:

$$z = \frac{(x-u)}{s} \quad (3)$$

Where z is values after transformation, u is the average value of the initial data, s is the standard deviation of the initial data and x is initial data.

The result of transformation steps P1 – P7 is the constructed forecast of the seasonal components and the residual in the form of noise and trend, which has yet to be predicted. To predict the remainder, the SARIMAX model is used, where additional information is considered as exogenous variables. Note that, in general terms, forecasting demand is reduced to determining the values of the matrix M or L in a certain horizon, however, the values of M or L available from the initial data reflect only the progress of sales and are distorted by restrictions on the capacity of $F1$, the impact of shares $F2$ and measures $F3$, competitors' prices $F4$. Before building a forecast, historical data are marked by dates. Each stream $M'_{i,j}$ for date d is assigned vectors of influence factors $F_i, j, d = (F1, F2, F3, F4)$, which are external variables for the SARIMAX learning algorithm.

In step P8, influence factors are generated. In the SARIMAX model, additional information from the production calendar is used as external variables:

- Day of the week.
- Month.
- Day of the month.
- Work / day off.
- Length of consecutive days off.

All of this information has been converted into attributes (predictors, independent variables):

- Average for the day of the week.
- Monthly average.
- Average for the day of the month.
- Work / day off.
- Length of consecutive days off.

The presence of promotions and events, as well as the factor influencing the prices of competitors are also external input parameters of the SARIMAX algorithm. Therefore, at this step, the influence factors $F2, F3, F4$ are generated. In step P9, another external variable is calculated, considering the influence factor of the capacity limitation.

In step P10, the parameters are selected for the SARIMAX (p, d, q) (P, D, Q, s) model. Here the parameters (P, D, Q, s) are responsible for modeling seasonal effects, and (p, d, q) - non-seasonal ones [15]. Based on the presetting procedure outlined above, it is supposed to fix the seasonality parameter (s), the number of times of seasonal differentiation (D) and the number of times of simple differentiation equal to the following values:

$$s = 7, D = 1, d = 0$$

Next, it is necessary to select the autoregressive parameters $AR (P, p)$ and the parameters of the moving average $MA (Q, q)$, where p is the order of the autoregressive process in the non-seasonal part of the model, P is the order of the autoregressive process in the seasonal part of the model, q is the order of the moving average process in the non-seasonal part of the model, Q is the order of the moving average process of the seasonal part of the model. For this, the search for the best parameters of the SARIMAX model was carried out by the method of full enumeration over the grid of parameters, in which all combinations of parameters from the following ranges were checked, determined based on the correlation functions:

$$p \in [1; 14], q \in [0; 14], P \in [0; 4], Q \in \{0; 1\}$$

Sampling these ranges yields 2250 parameter combinations. Information criteria are used to identify the best set. Here we used the most popular criteria AIC [16] and BIC [17]:

$$AIC = -2\ln(L) + 2k \quad (4)$$

$$BIC = k \ln(n) - 2\ln(L) \quad (5)$$

Where n is the number of observations, L is the maximum value of the likelihood function of the model, k is the number of its parameters.

3. Results and discussion

Figure 3 shows the result of forecasting for one of the passenger transportation markets. The discrepancy between Fact Test and Forecast allows you to judge the accuracy of the model. For this segment, the parameters ((14, 0, 14), (1, 0, 1, 7)) turned out to be the best on test data, and the MAPE error was 8%.

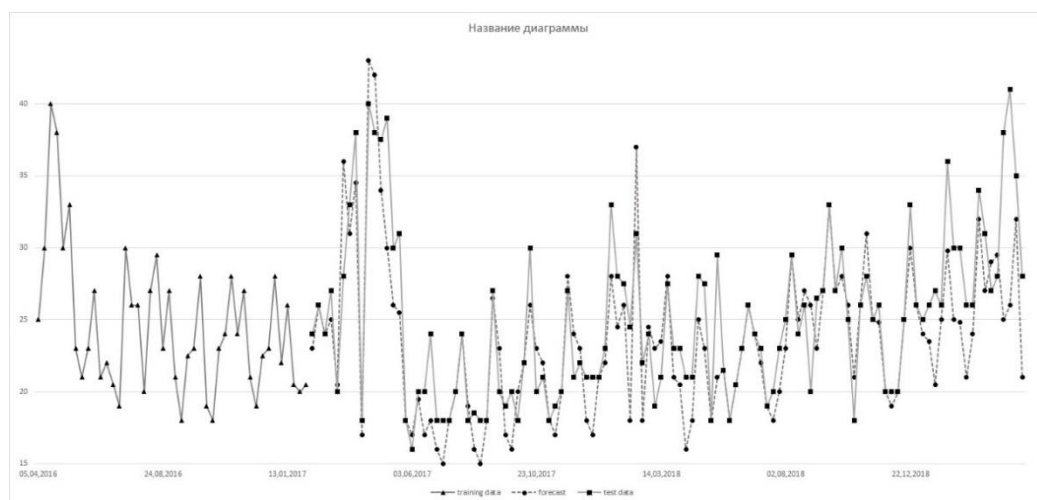


Figure 3. The result of selecting the optimal parameters for the test segment.

The article proposes an adaptive algorithm for forecasting the demand for passenger transportation, considering seasonality and external factors, such as limiting the capacity of the transport system, promotions, events and prices of competitors. To select a suitable predictive model, SARIMAX models with different parameters were tested. The test results show that SARIMAX ((14, 0, 14), (1, 0, 1, 7)) is most suitable for modeling demand in the selected rail passenger market.

The proposed approach to forecasting is flexible, which makes it capable of learning the patterns observed in various sections of passenger traffic. In this case, the best parameters of the model are selected automatically from the condition of maximizing the information criterion. This, in turn, means that the resulting vector of parameters of the SARIMAX model will differ from series to series. This is not a drawback of the model, but a direct consequence of its main advantage - adaptability to any fragment of the transport system. The created model can be used to support decision-making in the marketing and operational sectors of a transport company both independently and in combination with optimization algorithms.

Acknowledgment

The article was prepared with the financial support of the Center of Competence «Systems Engineering» established within the International Scientific and Methodological Center for the Transfer of Competencies of the Digital Economy of Ural Federal University, created in accordance with the Agreement dated 09.12.2019 No. 075-15-2019-1907

References

- [1] Smith B and Demetsky M 1997 Traffic flow forecasting: comparison of modeling approaches *J Transp Eng.* **123** 261-6

- [2] Do L, Taherifar N and Vu H 2019 Survey of neural network-based models for short-term traffic state prediction *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9** e1285
- [3] Zhang W, Yu Y, Qi Y, Shu F and Wang Y 2019 Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning *Transp A Transp Sci.* **15** 1688-711
- [4] Bernas M, Płaczek B, Porwik P and Pamuła T 2015 Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction *IET Intell Transp Syst.* **9** 264-74
- [5] Sun B, Cheng W, Goswami P and Bai G 2018 Short-term traffic forecasting using self-adjusting k-nearest neighbours IET Intell Transp Syst *Institution of Engineering and Technology* **12** 41-8
- [6] Castro-Neto M, Jeong Y, Jeong M and Han L 2009 Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions *Expert Syst Appl.* **36** 6164-73
- [7] Sun Y, Leng B and Guan W 2015 A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system *Neurocomputing* **166** 109-21
- [8] Hunt U 2003 Forecasting of railway freight volume: approach of estonian railway to arise efficiency *Transport* **18** 255-8
- [9] Chang Y and Liao M 2010 A seasonal ARIMA model of tourism forecasting: The case of Taiwan *Asia Pacific J Tour. Res* **15** 215-21
- [10] Milenkovic M, Svadlenka L, Melichar V, Bojovic N and Avramovic Z 2018 SARIMA modelling approach for railway passenger flow forecasting *Transport* **33** 1113-20
- [11] Okutani I, Stephanedes Y, Okutani I and Stephanedes Y 1984 Dynamic prediction of traffic volume through Kalman filtering theory *Transp Res Part B Methodol* **18** 1-11
- [12] Tang Y, Lam W and Pan N 2003 Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong *J Transp Eng* **129** 271-7
- [13] Box G, Box G and Cox D 1964 An analysis of transformations *Journal of the Royal Statistical Society* **B** 211-52
- [14] Cleveland R, Cleveland W, McRae J and Terpenning I 1990 STL: a seasonal-trend decomposition procedure based on loess *Journal of Official Statistics* **6** 3-73
- [15] Hyndman R and Athanasopoulos G 2018 Forecasting: Principles and Practice *Journal of computer and communications* **7** 421-55
- [16] Akaike H 1974 A New Look at the Statistical Model Identification *IEEE Trans Automat Contr* **19** 716-23
- [17] Schwarz G 1978 Estimating the Dimension of a Model *Annals of Statistics* **6** 461-4